1. A method to fabricate a twin MONOS memory comprising:

forming a deep N-well in a substrate;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer;

forming an oxide mask on sidewalls of said patterned cap nitride layer;

thereafter etching through said first polysilicon layer not covered by said cap nitride layer and said oxide mask to form first trenches and removing said ONO layer exposed within said first trenches by said etching;

forming oxide spacers on sidewalls of said first trenches;

thereafter depositing a second polysilicon layer within said first trenches and recessing said second polysilicon layer below said cap nitride layer;

depositing an oxide layer overlying said recessed second polysilicon layer wherein said recessed second polysilicon layer forms raised diffusions;

etching away remaining said first polysilicon layer not covered by said oxide mask leaving control gates underlying said oxide mask wherein said ONO layer lies only underneath said control gates and leaving second trenches;

depositing a second oxide layer lining said second trenches;

thereafter depositing a third polysilicon layer within said second trenches to form word gates between each two control gates and to form word lines overlying and crossing said word gates;

forming source and drain regions;

covering said gates with a dielectric layer; and

making contacts through said dielectric layer to said source and drain regions to form word line contacts and diffusion contacts to complete said Twin MONOS memory device.

2. The method according to Claim 1 wherein said step of forming said oxide-nitride-oxide (ONO) layer comprises:

thermally growing a base oxide layer on said substrate surface;

nitridizing said base oxide layer in a NH_3 ambient;

depositing a nitride layer overlying said nitridized base oxide layer;

depositing a top oxide layer overlying said nitride layer; and

oxidizing said nitride layer to stabilize a boundary between about nitride layer and said top oxide layer .

3. The method according to Claim 2 further comprising annealing said nitride layer in NO after depositing said nitride layer.

4. The method according to Claim 2 further comprising annealing said ONO layer in NH_3 or N_2O.

5. The method according to Claim 1 wherein said step of forming said oxide-nitride-oxide (ONO) layer comprises:

thermally growing a base oxide layer on said substrate surface;

nitridizing said base oxide layer in a NH_3 ambient;

depositing a nitride layer overlying said base oxide layer; and

oxidizing said nitride layer to form a top oxide layer overlying said nitride layer.

6. The method according to Claim 5 wherein said nitride layer is a silicon-rich nitride layer.

7. The method according to Claim 5 further comprising annealing said ONO layer in NH_3 or N_2O.

8. The method according to Claim 1 further comprising annealing said twin MONOS memory in H_2 after contact open process

9. The method according to Claim 1 wherein when electrons stored in a nitride portion of said ONO layer are to be erased through a bottom oxide portion of said ONO layer, a bottom oxide portion is thinner than a top oxide portion of said ONO layer.

10. The method according to Claim 1 wherein when electrons stored in a nitride portion of said ONO layer are to be erased through a top oxide portion of said ONO layer, a top oxide portion is thinner than a bottom oxide portion of said ONO layer.

11. The method according to Claim 1 wherein said oxide mask has a thickness of between about 20 and 80 nm.

12. The method according to Claim 1 further comprising implanting p-type dopant into said first polysilicon layer to eliminate an electron source through said ONO layer during F-N erase operation.

13. The method according to Claim 1 further comprising implanting boron or BF_2 ions at a tilt angle into said substrate under said first polysilicon layer to form a control gate channel.

14. The method according to Claim 13 further comprising implanting lightly doped n-type dopant into said control gate channel to prevent hole accumulation during F-N erase operation.

15. The method according to Claim 1 after said step of removing said ONO layer exposed within said trenches, further comprising implanting ions to form LDD regions.

16. The method according to Claim 1 wherein said raised diffusions form bit diffusions between said control gates to reduce bit resistance.

17. The method according to Claim 1 wherein said second polysilicon layer is recessed 50 to 150 nm from a top surface of said cap nitride layer.

18. The method according to Claim 1 wherein said word gate is a step word gate wherein said substrate is etched into a short distance during said step of forming said second trenches.

19. The method according to Claim 1 wherein said second trenches have a positive slope.

20. The method according to Claim 1 further comprising saliciding word lines.

21. A method to fabricate a twin MONOS memory in a CMOS process comprising:

providing a memory area and a CMOS area separated by isolation regions in a substrate;

forming a deep N-well in said memory area;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer in said memory area;

forming an oxide mask on sidewalls of said patterned cap nitride layer in said memory area;

thereafter etching through said first polysilicon layer not covered by said cap nitride and said oxide mask to form trenches and removing said ONO layer exposed within said trenches by said etching in said memory area;

forming oxide spacers on sidewalls of said trenches;

thereafter depositing a second polysilicon layer within said trenches and recessing said second polysilicon layer below said cap nitride layer;

depositing an oxide layer overlying said recessed second polysilicon layer wherein said recessed second polysilicon layer forms raised diffusions;

etching away remaining said first polysilicon layer not covered by said oxide mask leaving control gates underlying said oxide mask wherein said ONO layer lies only underneath said control gates and leaving second trenches;

patterning said first polysilicon layer in said CMOS area to form logic gates;

depositing a second oxide layer lining said second trenches and covering said logic gates;

thereafter depositing a third polysilicon layer in said memory area within said second trenches to form word gates between each two control gates and to form word lines overlying and crossing said word gates;

forming source and drain regions in said CMOS area and in said memory area;

covering said logic gates and said gates in said memory area with a dielectric layer; and

making contacts through said dielectric layer to said source and drain regions to form word line contacts and diffusion contacts to complete said Twin MONOS memory device.

22. The method according to Claim 21 wherein said step of forming said oxide-nitride-oxide (ONO) layer comprises:

thermally growing a base oxide layer on said substrate surface;

nitridizing said base oxide layer in a NH_3 ambient;

depositing a nitride layer overlying said nitridized base oxide layer;

depositing a top oxide layer overlying said nitride layer; and

oxidizing said nitride layer to stabilize a boundary between about nitride layer and said top oxide layer .

23. The method according to Claim 21 wherein said step of forming said oxide-nitride-oxide (ONO) layer comprises:

thermally growing a base oxide layer on said substrate surface;

nitridizing said base oxide layer in a NH_3 ambient;

depositing a nitride layer overlying said base oxide layer; and

oxidizing said nitride layer to form a top oxide layer overlying said nitride layer.

24. The method according to Claim 21 wherein when electrons stored in a nitride portion of said ONO layer are to be erased through a bottom oxide portion of said ONO layer, said bottom said oxide portion is thinner than a top oxide portion of said ONO layer.

25. The method according to Claim 21 wherein when electrons stored in a nitride portion of said ONO layer are to be erased through a top oxide portion of said ONO layer, a top oxide portion is thinner than a bottom oxide portion of said ONO layer.

26. The method according to Claim 21 wherein said oxide mask has a thickness of between about 20 and 80 nm.

27. The method according to Claim 21 further comprising implanting p-type dopant into said first polysilicon layer to eliminate an electron source through said ONO layer during F-N erase operation.

28. The method according to Claim 21 further comprising implanting boron or BF_2 ions at a tilt angle into said substrate under said first polysilicon layer to form a control gate channel.

29. The method according to Claim 21 further comprising implanting lightly doped n-type dopant into said control gate channel to prevent hole accumulation during F-N erase operation.

30. The method according to Claim 21 wherein said second polysilicon layer is recessed 50 to 150nm from a top surface of said cap nitride layer.

31. The method according to Claim 21 wherein said word gate is a step word gate wherein said substrate is etched into a short distance during said step of forming said second trenches.

32. The method according to Claim 21 after said step of removing said ONO layer exposed within said first trenches, further comprising implanting ions to form LDD regions.

33. The method according to Claim 21 further comprising saliciding said word lines.

34. A method to fabricate a twin MONOS memory in a CMOS process comprising:

providing a memory area and a CMOS area separated by isolation regions in a substrate;

forming a deep N-well in said memory area;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer in said memory area;

forming an oxide mask on sidewalls of said patterned cap nitride layer in said memory area;

thereafter etching through said first polysilicon layer not covered by said cap nitride and said oxide mask to form trenches and removing said ONO layer exposed within said trenches by said etching in said memory area;

forming oxide spacers on sidewalls of said trenches;

thereafter depositing a second polysilicon layer within said trenches and recessing said second polysilicon layer below said cap nitride layer;

depositing an oxide layer overlying said recessed second polysilicon layer wherein said recessed second polysilicon layer forms raised diffusions;

etching away remaining said first polysilicon layer not covered by said oxide mask leaving control gates underlying said oxide mask wherein said ONO layer lies only underneath said control gates and leaving second trenches;

depositing a second oxide layer lining said second trenches;

thereafter depositing a third polysilicon layer in said memory area within said second trenches to form word gates between each two control gates and to form word lines overlying and crossing said word gates and patterning said third polysilicon layer in said CMOS area to form logic gates;

forming source and drain regions in said CMOS area and in said memory area;

covering said logic gates and said gates in said memory area with a dielectric layer; and

making contacts through said dielectric layer to said source and drain regions to form word line contacts and diffusion contacts to complete said Twin MONOS memory device.

35. A twin MONOS memory comprising:

a deep N-well in a substrate;

polysilicon word gates having polysilicon control gates on sidewalls of said word gates having an oxide layer therebetween and an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

raised polysilicon diffusions between each two of said control gates and separated from said control gates by an oxide layer;

word lines overlying and crossing said word gates;

source and drain regions in said substrate underlying said raised diffusions;

diffusion contacts through a dielectric layer overlying said gates to said raised diffusions at an end of every other raised diffusion alternating in a memory block;

control gate contacts through a dielectric layer overlying said gates on an extension of said diffusion contacts or between diffusion contacts; and

word gate contacts at ends of said word lines alternately to complete said Twin MONOS memory device.

36. The device according to Claim 35 wherein said control gates have a thickness of between about 20 and 80 nm.

37. The device according to Claim 35 wherein said oxide-nitride-oxide (ONO) layer comprises:

a base oxide layer having a thickness of between about 3 and 6 nm;

a nitride layer overlying said base oxide layer having a thickness of between about 3 and 6 nm; and

a top oxide layer overlying said nitride layer having a thickness of between about 4 and 7 nm.

38. The device according to Claim 35 wherein electrons stored in said nitride layer are to be erased through said base oxide layer and wherein said base oxide layer is thinner than said top oxide layer.

39. The device according to Claim 35 wherein electrons stored in said nitride layer are to be erased through said top oxide layer and wherein said top oxide layer is thinner than said base oxide layer.

40. The device according to Claim 35 further comprising salicided said word lines.

41. A twin MONOS memory embedded in a CMOS device comprising:

a memory area and a CMOS area separated by isolation regions in a substrate;

a deep N-well in said memory area;

polysilicon word gates having polysilicon control gates on sidewalls of said word gates having an oxide layer therebetween and an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage in said memory area;

raised polysilicon diffusions in said memory area between each two of said control gates and separated from said control gates by an oxide layer;

word lines overlying and crossing said word gates;

polysilicon CMOS gates in said CMOS area;

source and drain regions in said substrate adjacent to said gates in said CMOS area and in said memory area;

diffusion contacts through a dielectric layer overlying said gates to said raised diffusions at an end of every other raised diffusion alternating in a memory block;

control gate contacts through a dielectric layer overlying said gates on an extension of said diffusion contacts or between diffusion contacts; and

word gate contacts at ends of said word lines alternately to complete said Twin MONOS memory device.

42. The device according to Claim 41 wherein said polysilicon gates have a thickness of between about 100 and 250 nm.

43. The device according to Claim 41 wherein said oxide-nitride-oxide (ONO) layer comprises:

a base oxide layer having a thickness of between about 3 and 6 nm;

a nitride layer overlying said base oxide layer having a thickness of between about 3 and 6 nm; and

a top oxide layer overlying said nitride layer having a thickness of between about 4 and 7 nm.

44. The device according to Claim 43 wherein electrons stored in said nitride layer are to be erased through said base oxide layer and wherein said base oxide layer is thinner than said top oxide layer.

45. The device according to Claim 43 wherein electrons stored in said nitride layer are to be erased through said top oxide layer and wherein said top oxide layer is thinner than said base oxide layer.

46. The device according to Claim 41 further comprising salicided said memory gates, said CMOS gates, and said source and drain regions in said CMOS areas.

47. A twin MONOS memory comprising:

a deep N-well in a substrate;

polysilicon word gates having polysilicon control gates on sidewalls of said word gates having an oxide layer therebetween and an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage and wherein said word gates form word lines;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate;

metal bit lines crossing said word gates and connecting to said local wiring; and

word gate contacts at ends of said word lines alternately to complete said Twin MONOS memory device.

48. A method to fabricate a twin MONOS memory comprising:

forming a deep N-well in a substrate;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer;

forming an oxide mask on sidewalls of said patterned cap nitride layer;

thereafter etching through said first polysilicon layer not covered by said cap nitride layer and said oxide mask to form first trenches and removing said ONO layer exposed within said first trenches by said etching;

forming oxide spacers on sidewalls of said first trenches;

thereafter depositing a second polysilicon layer within said first trenches and recessing said second polysilicon layer below said cap nitride layer;

depositing an oxide layer overlying said recessed second polysilicon layer wherein said recessed second polysilicon layer forms word gates;

etching away said cap nitride, underlying said first polysilicon layer not covered by said oxide mask, and underlying ONO layer, leaving control gates underlying said oxide mask on sidewalls of said word gates wherein said ONO layer lies only underneath said control gates and leaving second trenches;

implanting memory diffusions into said substrate underlying said second trenches;

thereafter forming a local wiring within said second trenches;

thereafter forming bit contacts to connect adjacent every other pair of memory diffusions through said local wiring underlying said bit contacts and through a metal line overlying said bit contacts and crossing said word gates and said local wiring; and

forming word gate contacts and control gate contacts at the ends of said word gates to complete said Twin MONOS memory device.

49. A twin MONOS memory comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate; and

metal bit lines crossing said control gates and isolated by shallow trench isolation (STI) lines to complete said Twin MONOS memory device.

50. The device according to Claim 49 wherein a width of each of said control gates and underlying said ONO is between about 30 and 60 nanometers.

51. A method to fabricate a twin MONOS memory comprising:

forming a deep N-well in a substrate;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer;

forming an oxide mask on sidewalls of said patterned cap nitride layer;

etching away said cap nitride leaving a looped said oxide mask on said first polysilicon layer;

selectively cutting said looped oxide mask at both ends of a block into two lines wherein adjacent control lines of adjacent loops are cut alternately;

thereafter etching away said first polysilicon layer not covered by said oxide mask and said underlying ONO layer, leaving control gates underlying said oxide mask wherein said ONO layer lies only underneath said control gates and leaving trenches;

implanting source/drain diffusions into said substrate underlying said trenches;

thereafter filling said trenches with oxide;

thereafter forming a common source line and bit contacts to some of said source/drain diffusions; and

forming bit line contacts and control gate contacts and connecting said bit contacts along said bit line by an overlying metal line to complete said Twin MONOS memory device.

52. An operation method for a twin MONOS memory comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate; and

metal bit lines crossing said control gates and isolated by shallow trench isolation (STI) lines to complete said Twin MONOS memory device

wherein said operation method comprises programming by electron ejection and erasing by electron injection comprising:

programming a selected memory cell by applying a high voltage to said control gate and a low voltage on a channel region underlying said control gate to eject electrons from said nitride portion of said ONO layer underlying said control gate through said top oxide portion of said ONO layer underlying said control gate; and

erasing all memory cells by injecting electrons into said nitride portion of said ONO layer through said top oxide portion of said ONO layer.

53. An operation method for a twin MONOS memory comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate; and

metal bit lines crossing said control gates and isolated by shallow trench isolation (STI) lines to complete said Twin MONOS memory device

wherein said operation method comprises programming by electron injection and erasing by electron ejection comprising:

programming a selected memory cell by applying a low voltage to said control gate and a high voltage on a channel region underlying said control gate to inject electrons into said nitride portion of said ONO layer underlying said control gate through

said top oxide portion of said ONO layer underlying said control gate; and

erasing all memory cells by ejecting electrons from said nitride portion of said ONO layer through said top oxide portion of said ONO layer.

54. An operation method for a twin MONOS memory comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate; and

metal bit lines crossing said control gates and isolated by shallow trench isolation (STI) lines to complete said Twin MONOS memory device

wherein said operation method comprises multi-level programming comprising:

controlling a threshold voltage of a selected memory cell by adjusting a voltage of said control gate or said bit line.

55. A twin MONOS memory comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

memory diffusions within said substrate between each two of said control gates wherein said memory diffusions are connected to adjacent diffusions alternately by a local wiring overlying said substrate;

metal bit lines crossing said control gates and isolated by shallow trench isolation (STI) lines wherein memory diffusions on one side of each of said control gates are connected to said bit lines through bit contacts; and

common source lines formed by connecting said memory diffusions on another side of each of said control gates with a local contact to complete said Twin MONOS memory device.

56. The device according to Claim 55 wherein a width of each of said control gates and underlying said ONO is between about 30 and 60 nanometers.

57. A method to fabricate a twin MONOS memory comprising:

forming a deep N-well in a substrate;

forming an oxide-nitride-oxide (ONO) layer overlying said substrate;

depositing a first polysilicon layer overlying said ONO layer;

depositing a cap nitride layer overlying said first polysilicon layer;

patterning said cap nitride layer to said first polysilicon layer;

forming an oxide mask on sidewalls of said patterned cap nitride layer;

etching away said cap nitride leaving a looped said oxide mask on said first polysilicon layer;

selectively cutting said looped oxide mask at both ends of a block into two lines wherein adjacent control lines of adjacent loops are cut alternately;

thereafter etching away said first polysilicon layer not covered by said oxide mask and said underlying ONO layer, leaving control gates underlying said oxide mask wherein said ONO layer lies only underneath said control gates and leaving trenches;

implanting source/drain diffusions into said substrate underlying said trenches;

thereafter filling said trenches with oxide;

thereafter forming local contacts to said source diffusions to form common source lines;

forming bit contacts to said drain diffusions and connecting said bit contacts along said bit line by an overlying metal line; and

forming control gate contacts to complete said Twin MONOS memory device.

58. A single gate MONOS memory device comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage; and

memory diffusions within said substrate between each two of said control gates wherein every other of said memory diffusions are connected to an overlying bit line through a local wiring overlying said substrate to complete said single gate MONOS device.

59. A method of programming a single gate MONOS memory device comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage; and

memory diffusions within said substrate between each two of said control gates wherein every other of said memory diffusions are connected to an overlying bit line through a local wiring overlying said substrate to complete said single gate MONOS device

wherein said method of programming comprises injecting electrons into said nitride portion of said ONO layer with channel hot electrons to store said electrons in each side of said control gate independently.

60. A method of erasing a single gate MONOS memory device comprising:

polysilicon control gates on a substrate having an oxide-nitride-oxide (ONO) layer underlying said control gates wherein portions of said substrate underlying said control gates define channels and wherein said nitride portion of said ONO layer underlying said control gates provides memory storage; and

memory diffusions within said substrate between each two of said control gates wherein every other of said memory diffusions are connected to an overlying bit line through a local wiring overlying said substrate to complete said single gate MONOS device

wherein said method of erasing comprises hot hole injection or Fowler-Nordheim ejection.

61. The method according to Claim 60 wherein hot holes are injected into both sides of said channels.

62. A method of device operation of a twin MONOS memory comprising:

polysilicon word gates having polysilicon control gates on sidewalls of said word gates having an L-shaped oxide-nitride-oxide (ONO) layer between said polysilicon word gates and said polysilicon control gates and underlying said control gates wherein said nitride portion of said ONO layer underlying said control gates provides memory storage;

wherein said method of device operation comprises:

performing one hot hole erasure after every n cycles of channel hot electron (CHE) programming and Fowler-Nordheim (F-N) erasure wherein electrons injected by said CHE programming are not perfectly erased by said F-N erasure and remaining said electrons are accumulated with said cycles and wherein said hot holes are generated by band to band transition by applying a negative bias on said word gate and are injected into said nitride portion of said ONO layer not only directly under one of said control gates but also at a corner of said L-shaped ONO layer wherein said hot holes neutralize accumulated said electrons within said nitride portion of said ONO layer at said corner of said L-shaped ONO layer.